A Survey of Machine Unlearning

Year: 2022 | Citations: 301 | Authors: T. Nguyen, T. Huynh, Phi-Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin

---

Abstract

Today, computer systems hold large amounts of personal data. Yet while such an abundance of data allows breakthroughs in AI, and especially machine learning, its existence can be a threat to user privacy, and it can weaken the bonds of trust between humans and AI. Recent regulations now require that, on request, private information about a user must be removed both from computer systems and from machine learning models—this legislation is more colloquially called “the right to be forgotten.” While removing data from back-end databases should be straightforward, it is not sufficient in the AI context as machine learning models often “remember” the old data. Contemporary adversarial attacks on trained models have proven that we can learn whether an instance or an attribute belonged to the training data. This phenomenon calls for a new paradigm, namely machine unlearning, to make machine learning models forget about particular data. It turns out that recent works on machine unlearning have not been able to completely solve the problem due to the lack of common frameworks and resources. Therefore, this article aspires to present a comprehensive examination of machine unlearning’s concepts, designs, methods, and applications. Specifically, as a category collection of cutting-edge studies, the intention behind this article is to serve as a comprehensive resource for researchers and practitioners seeking an introduction to machine unlearning and its formulations, design criteria, removal requests, algorithms, and applications. In addition, we aim to highlight the key findings, current trends, and new research areas that have not yet featured the use of machine unlearning but could benefit greatly from it. We hope that this survey serves as a valuable resource for machine learning researchers and those seeking to innovate privacy technologies. Our resources are publicly available at https://github.com/tamlhp/awesome-machine-unlearning.